UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

February 28, 2001

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES B-18*

MEMORANDUM FOR Howard Hogan
Chief, Decennial Statistical Studies Division

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Alfredo Navarro AN and Douglas Olson DO
Long Form and Variance Estimation Staff

Subject: Accuracy and Coverage Evaluation: Effect of Targeted Extended Search

The attached report was prepared, per your request, to assist the Executive Steering Committee on A.C.E. Policy in assessing the data with and without statistical correction.

The report focuses on the effects of the Targeted Extended Search (TES) operation and its effect on the components of the Dual System Estimates. The data contained herein are limited to those included in the A.C.E.

DSSD Census 2000 Procedures and Operations Memorandum Series B-18*
February 28, 2001

Accuracy and Coverage Evaluation: Effect of Targeted Extended Search

Alfredo Navarro and Douglas Olson
U.S. Census Bureau

Table of Contents

Tables

Accuracy and Coverage Evaluation: Effect of Targeted Extended Search

Prepared by Alfredo Navarro and Douglas Olson

Executive Summary

What is Targeted Extended Search (TES)?

The Targeted Extended Search (TES) operation is a relative small part of the Accuracy and Coverage Evaluation (A.C.E.), designed primarily to i) reduce variance associated with census geocoding error and ii) reduce bias from potential P-sample listing errors. The concept extended search, in which the blocks surrounding some A.C.E. block clusters are searched for matches and correct enumerations, is motivated by the fact that field workers are not 100 percent accurate when identifying whether a particular housing unit is inside or outside the cluster. Targeting means concentrating resources into clusters likely to produce the most payoff from searching the surrounding area. The TES is an operation by which the matching of households suspected to be erroneously geocoded in the census is extended into the surrounding area of some of the A.C.E. clusters.

Why perform TES?

Theoretically, it should be as likely for enumerators to mis-locate a housing unit outside its actual block as inside. As long as the errors of inclusion (counting a person living outside the cluster as inside) and exclusion (counting a person living inside the cluster as outside) are in balance, the net undercount as measured by the Dual System Estimates (DSEs) will not be affected. However, such persons would contribute either to lowered match and correct enumeration rates, which would make the DSE less precise by introducing additional variance. Since less than 10 percent of persons were non-matches or erroneous enumerations, an operation that allows 3 to 4 percent of that 10 percent to be match contributes significantly to lower variance.

The TES operation provides robustness against geocoding errors in the census and the A.C.E. For example, persons living in housing units inside the A.C.E. block clusters that are geocoded to a block in the surrounding area would contribute non-matches to the A.C.E. if an attempt was not made to search the surrounding areas.

What were the effects of TES?

For the nation as a whole in 2000, TES increased the Correct Enumeration rate by 2.9 percentage points and the Match Rate by 3.8 percentage points. Since there were more nonmatches than erroneous enumerations without TES, TES picked up more matches than erroneous enumerations in absolute numbers, but a smaller percentage.

Table I: Effect of TES on Correct Enumeration and Match Rates

	Correct Enumeration Rate (%)	Match Rate (%)
Without TES	92.3	87.7
With TES	95.3	91.6
Change from TES	+ 2.9	+3.8

Numbers may differ due to rounding.

What is "balancing error?"

Balancing error would exist if the treatment of the search area was inconsistent in the P- and E-samples. For instance, if the P-sample search area was defined as only the block cluster itself and the E-sample search area included neighboring blocks, then the expanded search area could increase the match rate (by allowing P-sample persons to match outside the search area) but not the Correct Enumeration rate, thus biasing the DSE's.

Was there balancing error in TES?

The much greater increase in the Match rate (3.8 percent) than the Correct Enumeration rate (2.9 percent) indicates that some aspect of A.C.E. may be out of balance or that TES is correcting potential P-sample geocoding error. The data do not exist at this time to determine if TES was the cause of that differential increase.

How do the results compare with the 1990 Census?

The corresponding operation in 1990 was called "Surrounding Block Search," which differed somewhat from the targeted operation in Census 2000. In 1990, the extended search increased the match rate by 4.1 percentage points and the correct enumeration rate by 2.3 percentage points. There is no particular reason to believe that the differences in results between the 1990 Post-Enumeration Survey and the 2000 A.C.E. resulted from methodological differences.

Background

The concept behind the Dual System Estimate is to estimate from a sample of block clusters the number of persons omitted from and erroneously enumerated in Census 2000. The complete definition of being omitted from or erroneously enumerated in the Census includes the concept of "location", that is, a successful enumeration must have located the person in the right place. Right location in the Census means anywhere in the block where the reported housing unit address was located, or in the "search area," defined as one ring of adjacent blocks. The operation concerned with counting and matching the persons in the surrounding areas is "Targeted Extended Search." The name was chosen because, unlike in the 1990 PES when the surrounding area of every cluster was searched, the A.C.E. search was "targeted" in two ways:

- Results from the Initial Housing Unit matching operation were used to select the housing units that are candidates for TES.

- In most cases, only clusters that include TES-eligible housing units were included in TES. The exceptions were clusters that were relisted, because their information was not available in time to be included in initial housing unit matching.

Using search areas, as long as a nonmatched census (E-sample) person resides in any of the blocks in the ring of surrounding blocks (and found not to be a census duplicate) he or she was labeled as a correct enumeration. Nonmatched persons in the P-sample found in the search area were treated as matches. If these two effects are balanced the measure of net undercount is not affected. Failure to balance the two effects results in "balancing error".

Balancing error can occur if the matching procedures describing the size of the search area are not applied consistently in the P- and E-samples. Differential effect on the correct enumeration and match rates can also be caused by geocoding errors in the P-sample. The latter would not be a balancing error in the TES, but in the A.C.E., and would be partly corrected by TES.

Balancing error in the A.C.E., if present, would lead to biases in the DSEs. Theoretically, DSE should be balanced as long as the geographic definition of the search area is consistent in the P- and E-sample, even if that search area was limited to just the cluster itself and no surrounding area. The purpose of TES is primarily variance reduction, but also to add robustness to the estimates. Variances will be low when the correct enumeration and match rates are consistent from cluster to cluster. TES helps make these rates more consistent by reducing the possibility of a person being counted as "wrong" because of being mis-located into a neighboring block. Robustness improves when TES corrects geocoding errors that are not otherwise taken into account. For instance, the determination of whether a census housing unit is a geocoding error is itself subject to geocoding error. Although there is no evaluation measure of potential geocoding error in the A.C.E. listing, operations were in place to minimize the chance of them occurring. For instance, A.C.E. clusters with a nonmatch rate of 45 percent or more were relisted and a geocoding check was performed on interviews where the interviewer changed the address. Consequently, these operations were designed to mainly prevent large levels of listing error.

The Effects of TES

What is the effect of TES at the national level?

The significance of TES is that it removes one of the sources of erroneous enumerations and nonmatches in the A.C.E., namely geocoding error. Regarding Table II, TES increased the number of Correct Enumerations from **244.4 million to 252.1 million** and Matches from **230.7 to 240.9 million**. Since over 90 percent of people match or are correctly enumerated, this helps reduce variance, which results principally from clusters that have large numbers of nonmatches or erroneous enumerations.

Table II: Effect of TES at the National Level

	With TES (1)	Without TES (2)	Difference (1) - (2)	Effect of TES (1) / (2)
E-sample				
Persons (Ne)	264,578,862	264,634,794	(55,932)	1.000
Correct Enumerations (CE)	252,096,238	244,387,951	7,708,288	1.032
CE Rate (%)	95.3	92.3	2.93	1.032
P-sample				
Persons (Np)	263,037,259	262,906,916	130,343	1.000
Matches	240,878,622	230,681,205	10,197,418	1.044
Match Rate (%)	91.6	87.7	3.83	1.044
Ratio of CE to Match Rate	1.040	1.053	(0.012)	0.989
Standard Error of Ratio (%)	0.134	0.331	(0.197)	40.5

Note: Table above reflects national totals without regard to post-stratification and will differ from other totals in which post-strata were summed.

What was the effect of TES by Regional Office?

TES made the Correct Enumeration (CE) and Match rates significantly more consistent across regional offices. Without TES, CE rates ranged from 90.0 percent to 94.6 percent; TES reduced that range to only 2.9 percent, from 93.3 to 96.2 percent. On the P-sample side, the range of Match rates was reduced from 7.7 percent (84.1 to 91.8) to 5.4 percent (88.7 to 94.1). The results are consistent with the hypothesis that TES helps to reduce large potential errors and inconsistencies in data collection.

Table III - Effect of TES by Regional Office

	E-sample CE Rate (percent)		P-sample Match Rate (percent)		Ratio CE Rate / Match Rate			
	With TES	No TES	With TES	No TES	With TES	No TES	Change (pct)	Stand Error
Boston	95.8	93.4	92.0	89.6	1.040	1.042	-0.2	0.50
New York	93.3	90.0	88.7	84.4	1.051	1.065	-1.4	1.78
Philadelphia	95.5	91.4	91.9	87.2	1.039	1.048	-0.9	0.73
Detroit	96.2	93.8	94.0	90.2	1.023	1.040	-1.7	0.73
Chicago	95.8	92.8	92.5	89.0	1.036	1.043	-0.7	0.71
Kansas City	96.2	94.6	94.1	91.8	1.022	1.031	-0.9	0.41
Seattle	95.0	92.6	91.4	87.7	1.039	1.055	-1.6	0.69
Charlotte	95.5	91.6	91.3	87.6	1.045	1.045	0.0	0.86
Atlanta	94.6	91.1	90.4	84.1	1.046	1.084	-3.8	1.77
Dallas	94.5	91.6	89.9	86.7	1.051	1.057	-0.6	0.82
Denver	95.0	92.9	91.4	88.6	1.039	1.048	-0.9	0.63
Los Angeles	95.8	92.0	91.1	86.6	1.052	1.062	-1.0	1.83

The consistency of data among these various groups suggests that the operations were performed consistently throughout the sample. The one large group that draws attention is the Atlanta Regional Office, whose CE/Match ratio was lowered by 3.8 percent, from 1.084 to 1.046, by TES. This office started with the lowest match rate of any Regional Office and TES brought it up to the normal range. The difference between its change in adjusted rate ratio (3.8 percent) and the national average change (1.2 percent) is less than what would be considered statistically significant, given its standard error of 1.77 percent.

What were the effects of TES for sub-population groups?

The effect of TES is also very consistent among large population sub-sets, such as age/sex groups and race groups. In some of the smaller race groups, like American Indians and Pacific Islanders, the patterns are not as consistent, as might be anticipated because of the larger relative variance of smaller groups and the listing difficulties in the areas in which they live. **TES affected the seven Age/Sex domains about equally, since there is no reason for those variables to correlate with listing difficulty.**

Table IV -- Effect of TES on Race/Origin Domains

Race/Origin Domain	E-sample CE Rate (percent)		P-sample Match Rate (percent)		Ratio CE Rate / Match Rate			
	With TES	No TES	With TES	No TES	With TES	No TES	Chg (percent)	Std. Err. Chg.
Am Ind on Reserv	95.81	93.24	85.99	77.80	1.114	1.198	-8.4	2.50
Am Ind off Reserv	93.97	91.85	87.54	84.77	1.073	1.083	-1.0	1.19
Hispanic	94.46	91.84	87.47	83.33	1.080	1.102	-2.2	0.91
Black	92.73	89.60	86.94	82.69	1.067	1.084	-1.7	0.76
Pacific Islander	93.05	91.01	84.66	81.67	1.099	1.114	-1.5	2.48
Asian	94.57	90.47	90.45	86.24	1.046	1.049	-0.3	1.45
White & Other	95.90	92.99	93.12	89.43	1.030	1.040	-1.0	0.25

Table V -- Effect of TES on Age/Sex Domains

AGE/SEX	E-sample CE Rate (percent)		P-sample Match Rate (percent)		Ratio CE Rate /Match Rate			
	With TES	No TES	With TES	No TES	With TES	No TES	Chg (percent)	Std. Err. Chg.
0-17	95.94	93.09	90.84	86.97	1.056	1.070	-1.4	0.40
18-29 M	92.87	89.61	86.40	82.42	1.075	1.087	-1.2	0.58
18-29 F	93.61	89.92	88.54	84.36	1.057	1.066	-0.9	0.71
30-29 M	95.23	92.28	91.23	87.53	1.044	1.054	-1.0	0.34
30-49 F	96.01	93.04	92.90	89.06	1.033	1.045	-1.1	0.32
50+ M	95.34	92.67	93.68	90.03	1.018	1.029	-1.2	0.27
50+ F	95.51	92.84	94.29	90.55	1.013	1.025	-1.2	0.31

The only population group for which the CE rate/Match rate ratio change was significantly different from the national average of 1.2 percent was American Indians on Reservations, their CE to Match rate changed 8.4 percent under TES. In 1990, this population group also saw its CE/Match ratio drop by 6.6 percent under TES, suggesting major listing problems in these geographic areas.

What was the effect of TES on Variance?

The table below shows the significant contribution that TES makes to variance reduction. For the A.C.E. considered as a whole (i.e. a direct DSE of the entire population) the coefficient of variation is 0.129 percent with TES and 0.314 percent without it. At the post-stratum level, the average weighted improvement is about 33 percent. The gains in precision as measured by variance are obvious. So there can be little question that TES makes DSE estimates more precise, and that TES improves the quality of the A.C.E. so long as it does not make DSE less accurate by introducing bias.

Table VI: Effect of TES on Coefficient of Variation

	Std. Err.	CV (percent)
TES Performed	355,451	0.129
Without TES	877,664	0.314

Reduction occurred in the CV of a majority of the collapsed post-strata (448 original post-strata were collapsed into 416 post-strata for DSE calculation purposes.)

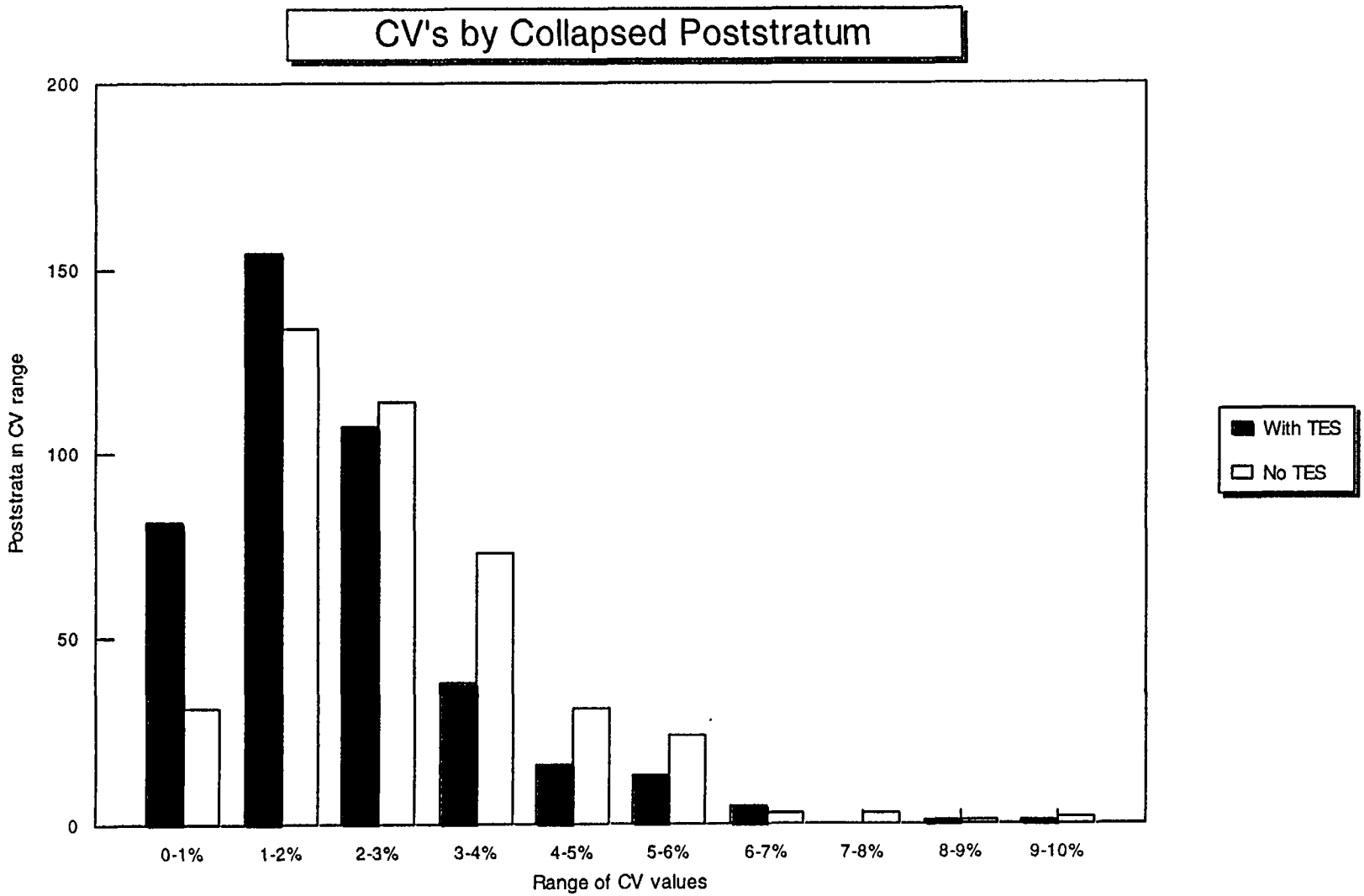
Table VII – Effect of TES on Post-stratum CV's (percents)

	With TES	No TES
Average CV	2.07	2.66
Median CV	1.81	2.32
Average CV weighted by Census Count	1.30	1.93

The average and median improvement in the CV, under the assumption that TES was not performed, was about 22 percent. A study using the 1990 Post-Enumeration Survey data concluded that performing TES decreased CV's by an average 20 percent. See [1]. Although the measurements are not strictly comparable due to differences in methodology, we would expect the variance increase to be approximately similar.

Chart I on the next page illustrates the effect TES had on the variances of the collapsed post-strata, showing clearly that greater number of post-strata with very small CV's (2 percent or less) and the smaller number of post-strata with CV's larger than 5 percent.

Chart I – Effect of TES on Post-stratum Variances



Was there any Balancing Error?

A severe limitation of this study is that no data is available to study the quality of the production data except the production data itself. The two instruments, the Census and the A.C.E., can help to serve as quality checks against each other and, in fact, that is the purpose behind the Dual System Estimate. Unfortunately, there are no additional evaluation survey available at this time to check the quality of the DSE's themselves. Therefore, the best measures possible will be developed using the P- and E-sample data themselves to check for the existence of balancing error in the A.C.E.

One of the ways that TES balance can be estimated is to assume that the number of errors of inclusion and exclusion should be approximately equal and hence that the number of TES people "found" on the P-sample and E-sample sides should be about equal. (Here "found" means to be a P-sample Match or E-sample Correct Enumeration.) In the 2000 A.C.E., it appears that about 3 million persons could be out of balance after adjusting for P-sample coverage:

Table VIII: P and E Sample Surrounding Block Matches

	Count*	Weighted
E-sample CE's in surrounding blocks	20,401	7,708,288
P-sample Matches in surrounding blocks	21,878	10,002,072
Adjusted for P-sample coverage		10,676,849

*May differ from other unweighted counts because of treatment of unresolved cases and treatment of certain persons in relist clusters who were not affected by TES

Lacking any data with which to measure the possible causes of this imbalance, we are left with proposing a couple of theories, and explaining whether TES more likely contributed to imbalance or helped to re-balance an inherently imbalanced system.

Possible Imbalance in Field Listing

One possible cause of imbalance could come from the field work done in the Initial Housing Unit Matching operation. Enumerators were instructed to differentiate between housing units that did not exist and those that were in the search area (i.e. the surrounding blocks.) Failure to accurately distinguish between the two cases could have resulted in some surrounding block housing units being treated as erroneous enumerations and hence not eligible for search. In fact, erroneously enumerated housing units did included almost 2 million E-sample correct enumerations, at least some of whom should probably have been included in TES.

Table IX: Erroneous Enumeration Housing Units

	Count	Weighted
Total EE Housing Units	5,996	1,724,645
With E-sample persons	3,450	1,039,254
No-E-sample persons	2,546	685,391
Persons in EE units	8,104	2,448,863
Correct Enumerations	6,439	1,924,233

A follow-up field evaluation, TES2, is currently underway as of this writing to determine the degree of confounding between EE and Geocode Error housing units. Once the results from the Evaluation Studies become available we will learn more about these errors. An area of interest is to analyze the results looking at the number of cases that shift from the A.C.E. cluster (GC) or "Surrounding Block" (GS) to Erroneous (GE) or vice-versa. These results may help to explain the so called A.C.E. imbalance.

The same problem that exists for Erroneous Enumeration housing units probably also exists for correct enumeration housing units. We know from the TES results that about 10 percent of persons in housing units coded as geocoding error were actually within the sample block (i.e. the geocode error determination itself was a geocode error.) We know from the handwritten notes of field workers that the opposite problem also occurs – housing units coded as correctly enumerated within the block were actually geocode errors. This error does not directly effect the DSE, with or without TES, but lowers the measured number of matches in surrounding blocks (such persons would have been recorded as correct enumerations instead) and hence contributes to an increase in the apparent measurement of the imbalance.

Geocoding error in the P-sample

Another source of imbalance is that the P-sample listing could have included some housing units in surrounding blocks. P-sample listing errors of exclusion (i.e. failing to list P-sample housing units inside the block cluster) would have little effect on DSE. However, P-sample listing errors of inclusion (listing as in the cluster housing units actually in the neighboring blocks) would introduce imbalance by including in the P-sample persons with no chance of matching if TES was not performed. TES helped to correct for this kind of imbalance by giving every whole household non-match that was also an address non-match a chance to be included in TES matching operations and to match to a housing unit in the surrounding block (Childers, 2001). A study is underway at this writing to attempt to quantify the extent of P-sample geocode errors. There was some evidence of P-sample geocoding error in the 1990 PES, with a national estimate of 434,000 P-sample people geocoded to outside the sample cluster. However, this level of error did not result in a measurable error in matching. An error is defined as a person who matches as

a result of redefining the search area. We believe that P- sample geocoding error in the 2000 A.C.E. is smaller than in the 1990 PES, because of the relisting of clusters with high nonmatch rates and quality control operation for the A.C.E. listing (Childers, 2001).

Operational change made during processing

After the TES was designed, it was discovered that a potential imbalance existed in the way added and deleted housing units were being handled. Since the TES housing unit inventory was determined by the Initial Housing Units Matching, changes to the housing unit inventory subsequent to that operation would not be reflected in the TES (Beaghen, 2000). To adjust for these changes, the TES processing procedures were modified to:

1) Perform a surrounding block search on any P-sample person in a housing unit that was matched to an E-sample unit that was subsequently deleted;

2) Impute a Correct Enumeration probability for any person in an added unit determined to be a geocoding error.

Discussion

In 1990, the corresponding Surrounding Block Search operation identified 9.9 million matches (4.1 percent of P-sample persons) and 5.6 million correct enumerations (2.3 percent of E-sample persons) in surrounding blocks, an apparent imbalance of 4.3 million weighted persons. A follow-up study to the 1990 Census did not conclude with certainty that balancing error was either present or absent (Bateman, 1991).

Although it is not possible to measure directly, it is reasonable to assume that geocoding error in the P-sample exists to some extent. The P-sample re-listing of clusters that showed a nonmatch rate of 45 percent or greater eliminated some P-sample geocoding error, but certainly not all. Since we cannot find any direct evidence of balancing error in any other part of the operation, P-sample geocoding error remains the most likely cause of the measured imbalance in the number of surrounding block matches.

Was the TES sample biased?

One possible, but unlikely, source of balancing error would have occurred if the actual sample selected for TES was out of balance with respect to the underlying population, in that a preponderance of E- or P-sample persons could have been included in the sample out of proportion the their numbers in the sampling universe. This was not, however, the case because the actual sample selected included approximately the same number of persons as in the universe once weights were applied.

Table X: Persons in TES Sample Before and After Sampling

	E-sample	P-sample
Count		
Universe	32,334	48,542
TES-Weighted Sample	31,911	48,464
Weighted		
Universe	8,786,027	15,313,665
TES-Weighted Sample	8,730,175	15,469,782

Limitations

The principal limitations of this study have been discussed in the body – lack of information outside the A.C.E. production data itself with which to conduct an evaluation of possible balancing error.

- A.C.E. Listing Geocode Error – The data simply do not exist to measure the possibility that housing unit and persons were not erroneously included or excluded from the A.C.E. listing through geographic errors. Doubtless these errors occurred to some extent. Some errors of inclusion were counted in TES, helping to offset a potential source of balancing error that would have occurred if TES had *not* been performed. There is anecdotal evidence from field work observation that housing units far outside the cluster boundaries have sometimes been included in the P-sample, but the scope of such errors is not measurable with available data.

- Distinction between Erroneous Enumerations and Geocode Error in the E-sample – The identification of E-sample TES-eligible housing units relied on a field identification of the location of the unit. To the extent that field workers weren't able to distinguish between housing units within and outside the search area (which should have been erroneous enumerations) the inventory of TES-eligible units on the E-sample side was incomplete. There is an additional field evaluation, called "TES2" that will visit many EE housing units to determine how many could have been more appropriately coded as geocode errors, but such data will not be available until that operation has been performed.

- Housing units erroneously coded as inside the cluster – Units that were coded as inside the cluster were not subject to additional field follow-up. We know from the follow-up of units geocoded outside the cluster that about 10 percent were coded as being inside, in other words that two different enumerators disagreed about the geographic location of the unit. It is possible that a complete follow-up of units coded inside the cluster might have determined that many were actually outside and should have been counted as TES housing units. The type of error described here would not direct effect Dual System Estimates but would help to explain the difference between the number of matches and correct enumerations found in surrounding blocks.

Conclusions

To the extent that its effects can be measured, TES contributed significantly to increasing the correct enumeration and match rates and hence to the reduction of variance in the A.C.E. Examination of detailed data suggests that TES's contributions were nearly uniform across regional areas and within demographic groups whose populations are large enough to limit the effects of variance. The observed differences between the number of matches and correct enumeration in surrounding blocks for some population groups and at the RO level are an indication that balancing error may have been present. However, based on the data we conclude that balancing error was not directly observed and that further evaluations would be necessary to determine conclusively if balancing error was or was not present.

References

1. Singh, Rajendra P., "Search Area Definition for the Dual System Estimation", October 31, 1997, memorandum for Elizabeth A. Vacca, U.S. Bureau of the Census.

2. Bateman, David V., "Post Enumeration Survey Evaluation Project P11: Balancing Error Evaluation", 1990 Coverage Studies and Evaluation Memorandum Series #M-2

3. Beaghen, Michael;"Accuracy and Coverage Evaluation (A.C.E.) Targeted Extended Search (TES) Modifications", Census 2000 Procedures and Operations Memorandum Series

4. Childers, Danny, "Accuracy and Coverage Evaluation: The Design Document," January 26, 2001, Census 2000 Procedures and Operations Memorandum Series, S-DT-01

5. Childers, Danny and Xijuan Liu, "Accuracy and Coverage Evaluation: Additional Geographic Coding for Erroneously Enumerated Housing Units", February 28, 2001, Census 2000 Procedures and Operations Memorandum Series, T-6

6. Olson, Douglas; "Accuracy and Coverage Evaluation Survey – Identification and Sampling of Block Clusters for Targeted Extended Search", October 22, 1999, Census 2000 Procedures and Operations Memorandum Series R-20

Appendix 1: TES in DSE Estimation

Applying the TES operation to the DSE calculation is primarily a matter of applying weights properly. Every person in the A.C.E. is either a TES person or a non-TES person and every A.C.E. cluster is either a TES cluster or a non-TES cluster. TES does not in any way effect the weights of non-TES persons. Every TES person is assigned the TES weight of his A.C.E. cluster, which is:

- Zero, if the cluster was not selected for TES. In practice, only persons in TES-eligible housing units that include clusters eligible for the sampling operation, but which were not selected for the sample, can fall into this category.
- One, if the cluster was selected for TES with certainty.
- 4.9, if the cluster was selected for TES in the systematic sampling phase of the TES cluster selection. Conceptually, the persons in the TES sample are weighted up to make up for the TES-eligible persons who are assigned a weight of zero because they were *not* selected to be in the sample.

The calculation of the DSE requires the use of seven components, each of which represents the sum of the A.C.E. weights for some group of persons in the A.C.E., including both TES and non-TES persons. Hence, each of the seven components represents a weighted sum of TES and non-TES persons, the latter with their TES cluster weights applied. TES weights are multiplied by the A.C.E. weight of the cluster for TES persons as in the following table:

	TES Cluster	non-TES Cluster
TES Persons	1, if cluster in TES with Certainty	zero
	1/(TES Sampling Rate), if cluster selected for TES by Sampling	zero
non-TES Persons	1	1

DSE Calculation

For Census 2000 the DSE estimator is:

$$\hat{DSE} = (DD)\left(\frac{CE}{N_e}\right)\left(\frac{N_n + N_i}{M_n + \left(\frac{M_o}{N_o}\right)N_i}\right)$$

where :

DD	= data-defined census persons, (excludes late adds)
II	= count of not-data-defined and wholly imputed persons
CE	= estimated number of A.C.E. E-Sample correct enumerations
N_e	= estimated number of A.C.E. E-Sample persons
N_n	= estimated number of A.C.E. P-Sample nonmovers
N_i	= estimated number of A.C.E. P-Sample inmovers
N_o	= estimated number of A.C.E. P-Sample outmovers
M_n	= estimated number of A.C.E. P-Sample nonmover matches
M_o	= estimated number of A.C.E. P-Sample outmover matches

The estimator has seven A.C.E. components (DD comes from the Census Initial Phase Enumeration and is not part of DSE). Each of the seven components represents a weighted sum of persons, which will include both TES and non-TES persons. TES persons can be in-movers, out-movers and non-movers just like all other persons.

Each of the seven DSE components is expressed as follows:

$$\sum_{i=1}^{n}\sum_{j=1}^{n_p} w_{ij}^{*}\, m_{ij}\, x_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n_p} w_{ij}^{*}\, m_{ij}\, y_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n_p} w_{ij}^{*}\, t_{ij}\, m_{ij}\, z_{ij}$$

where;
i= cluster index
j= person index
n= number of block clusters in the A.C.E. sample
x_{ij} = 1 if the person is not a TES person, 0 otherwise
y_{ij} = 1 if the person is a TES person and is in the TES sample with certainty, 0 otherwise
z_{ij} = 1 if the person is a TES person and is in the TES systematic sample, 0 otherwise
m_{ij} = characteristic of interest, match, correct enumeration, E-sample person, or P-sample person

w_{ij}*= weight used for estimation (includes inverse of the probability of selection for A.C.E., adjustment for household noninterview and missing data imputation)
t_{ij} = TES sampling weight, the TES systematic sample take-every

Appendix 2 – TES Sampling

The principles of TES sampling for Census 2000 were based on our experiences with the 1990 Census. That year, surrounding block search was conducted in every A. C.E. cluster (A.C.E. was called PES then). Most clusters, however, did not include any geocode errors so the surrounding block operations in those clusters was a waste of time and effort and might have led to errors by the matchers. Therefore, for Census 2000 it was decided that surrounding block search would be conducted only in a sample of A.C.E. clusters, in order to better target the operations into clusters in which they were necessary (Olson, 1999).

To this end, TES sampling was performed using the following principles:

- Every cluster containing a TES-eligible housing unit must have some probability of inclusion into TES.
- Every TES-eligible housing unit is assigned a TES-sampling weight equal to the reciprocal of its probability of inclusion. This ensures that every TES-eligible housing unit has an expected TES weight of 1.
- The A.C.E. clusters with the most TES-eligible housing units would be included in TES with certainty and assigned a TES weight of 1.

In order to achieve these goals, TES sampling followed the following steps:

- Include in TES with certainty the five percent of clusters with the most TES-eligible housing units.
- Include in TES with certainty the five percent of clusters with the most ACE-weighted TES-eligible housing units.
- Sample among all other A.C.E. clusters containing TES-eligible housing units.

These steps were performed sequentially without replacement of already-selected clusters.

By performing the sampling in this way, about 70 percent of all TES-eligible housing units were included in TES (60 percent were included with certainty and approximately one in five of the other 40 percent). Every TES-eligible housing unit still maintained an expected TES sampling weight of unity, and by including most of the TES-eligible housing units with certainty, the variance associated with TES sampling was kept to a minimum.

Because the information to perform TES sampling was not available for relist clusters at the time the sample was drawn, all such clusters were included in TES. List/Enumerate clusters were excluded from TES. The following table lists how many clusters and persons were included in TES under each of the phases of the sampling:

Size of TES Sample

		E-sample Persons		P-sample Persons	
	Clusters	Size	Weighted (000's)	Size	Weighted (000's)
Included with certainty	1,088	21,755	5,829	27,354	9,375
Sampled In	1,089	2,281	555	3,788	1,209
Sampled Out	4,237	8,298	2,217	14,741	4,566

Appendix 3 – Variance Estimation Assumptions

The variance estimation technique used in this report is a stratified jackknife on the clusters in the final A.C.E. sample of clusters. The strata represent A.C.E. reduction strata, which were used to select a final representative selection of clusters. The jackknife calculation employs a "delete one" methodology in which replicates are calculated under the assumption that one cluster has been deleted from the sample, and then the rest of the clusters in the same stratum are re-weighted as if the deleted cluster had not been part of the sample to being with.

To perform this calculation quickly, certain assumptions (some of the erroneous assumption employed for simplification purposes had to be employed):

- The weighting of TES persons representing TES sampling adequately adjusts for the variance contribution of TES sampling without further adjustment to the replicate weights.
- The variance contribution of strata that include only one cluster can be estimated by calculating the effect on DSE's as if there had not been such a strata/cluster.
- Imputed values for unresolved cases were not replicated and therefore the contribution of missing data error to the overall variance is not accounted for.
- The final sampling is treated as if it had been drawn by a one-stage sampling operation.

The variance of the DSE is calculated as follows:

$$Var(DSE) = \sum_{i}^{Clusters} ((n_{FSS} - 1) / n_{FSS}) x (DSE_i - DSE_0)^2$$

n_{FSS} = number of clusters in final sampling stratum

DSEi = DSE for the i^{th} replicate

DSE0 = DSE for the full sample